

Network Flow-based Simultaneous Retiming and Slack Budgeting for Low Power Design

Bei Yu*, Sheqin Dong*, Yuchun Ma*, Tao Lin*, Song Chen[†] and Satoshi Goto[†]

*Department of Computer Science & Technology, Tsinghua University, Beijing, China

[†]Graduate School of IPS, Waseda University, Kitakyushu, Japan

Email: {b-yu07@mails, dongsq@mail}.tsinghua.edu.cn

Abstract—Low power design has become one of the most significant requirements when CMOS technology entered the nanometer era. Therefore, timing budget is often performed to slow down as many components as possible so that timing slacks can be applied to reduce the power consumption while maintaining the performance of the whole design. Retiming is a procedure that involves the relocation of flip-flops (FFs) across logic gates to achieve faster clocking speed. In this paper we show that the retiming and slack budgeting problem can be formulated to a convex cost dual network flow problem. Both the theoretical analysis and experimental results show the efficiency of our approach which can not only reduce power consumption but also speedup previous work.

1. INTRODUCTION

Timing constraint design and low power design have become significant requirements when the CMOS technology entered the nanometer era. On the one hand, more and more devices trend to be put in the small silicon area while at the same time the clock frequency is pushed even higher. As an effective timing optimization scheme, retiming is a procedure that involves the relocation of flip-flops (FFs) across logic gates to achieve faster clocking period. On the other hand, to tackle the tremendous growth in the design complexity, timing budgeting is performed to relax the timing constraints for as many components as possible without violating the system's timing constraint. Therefore, both retiming and timing budget might influence the timing distribution of the design greatly.

Since Leiserson and Saxe proposed the idea of retiming in 1983 [1], it has become one of the most powerful sequential optimization techniques. In [2], the min-area retiming problem was solved by min-cost network flow algorithm. Recent publications [3] and [4] proposed a very efficient retiming algorithm for minimal period by algorithm derivation. [5] and [6] respectively presented efficient incremental algorithms for min-period retiming under setup and hold constraints, and min-area retiming under given clock period.

For timing-constrained gate-level synthesis, timing slack is an effective method for circuit's potential performance improvement. The components with relaxed timing constraints can be further optimized to improve system's area, power dissipation, or other design quality metrics. The slack budgeting problem has been studied well. Some of the previous slack budgeting approaches are suboptimal heuristics such as Zero-Slack Algorithm (ZSA) [7]. [8][9] formulated the slack budgeting problem as Maximum-Independent-Set (MIS) on

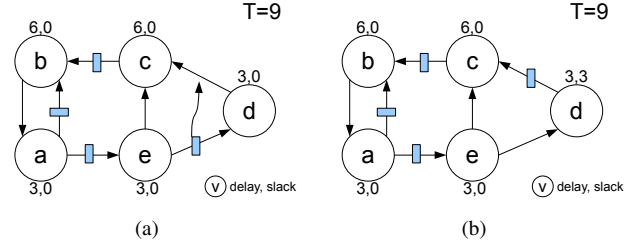


Fig. 1: Relocate FFs to increase potential slack without violating timing constraint. (a) No potential slack in this circuit. (b) moving the FF from edge de to edge cd , the potential slack can be increased from 0 to 3.

sensitive transitive closure graph. In [10] and [11], authors proposed combinatorial methods based on net flow approach to handle the slack budget problem.

Budgeting problem can be extended to describe exactly real-world applications, such as gate resizing, multiple V_{dd} and multiple V_{th} assignment [12][13][14]. Since the number of logically equivalent cells in a library is limited, it is reasonable to limit the possible slack value in real designs. In [15], Qiu et al. showed that power reduction is not proportional to the slack amount and propose a piecewise linear model to approximate the relationship between slack and power reduction. In this paper, we adopt the same model and consider discrete slack budgeting problem. Note that our method can be easily transferred into continuous slack budgeting problem.

Nearly all the existing slack budgeting algorithms are either used for combinatorial circuit, or limited to fixed FF locations. At the early design stages, it is flexibility to schedule pipeline or timing distribution to obtain more timing slack. As shown in Fig. (1), the period of a circuit is minimized with the delay and slack labeled beside each gate as well. It is seen that there is no potential slack in this circuit. However, if retiming and slack budget process is taken, *i.e.* moving the FF from edge de to edge cd , the potential slack can be increased from 0 to 3, keeping the period minimized at the same time.

A simultaneous retiming and slack budgeting algorithm for dual- V_{dd} programmable FPGA power reduction was proposed in [16]. In [17] Lin et al. proved that slack budgeting problem can be viewed as a convex retiming problem. However they

failed to formulate retiming and slack budgeting simultaneously. In [18] authors proposed a simultaneously slack budgeting and incremental retiming algorithm to maximize the potential slack by retiming for synchronous sequential circuit. They proposed a reasonable algorithm flow, however, their solution quality suffers in two aspects. First, there was no guarantee that the algorithm will get optimal solution because iterative strategy is easily trapped in local optimum. Besides, the slack budget problem was translated to a Maximal Independent Set (MIS) problem, which is a NP-hard problem.

[19] showed that for an Integer Linear Programming (ILP) with separable convex objective functions and special form of constraints, it can be viewed as convex cost dual network flow problem and solved in polynomial time. This model has been adopted in various works, such as buffer insertion [20], multi-voltage supply [21][22], clock skew scheduling [23] and slack budgeting [17].

In this paper we first formulate retiming and slack budgeting problem as an Integer Linear Programming (ILP) problem. Since ILP has been listed to be one of the known NP-hard problems, we then show how to transform this problem to the convex cost dual network flow problem with just a little loss of optimality. Experimental results show that our algorithm can not only reduce power consumption, increase total slack budgeting, but also effectively speedup previous work.

The remainder of this paper is organized as follows. Section 2 defines the simultaneous slack budget and retiming problem. Section 3 presents our algorithm flow. Section 4 reports our experimental results. At last, Section 5 concludes this paper.

2. PROBLEM FORMULATION

As shown in [1], we model a synchronous sequential circuit as a directed graph $G(V, E, d, w)$, each vertex $i \in V$ represents a combinational gate and each edge $(i, j) \in E$ represents a signal passing from gate i to j . Non-negative gate delays are given as vertex weights $d : V \rightarrow R$. Non-negative integer $w : E \rightarrow Z$ as the edge weight represents the number of FFs on the signal pass. The max clock period is given as T .

For each vertex, three non-negative labels, $a_i/\gamma_i/s_i$, represent the latest arrival time, require time, and slack of vertex i . a_i and γ_i can be calculated as follows:

$$\begin{cases} a_i = d_i & \text{if } w(k, i) > 0 \text{ or } i \in PI \\ a_i = \max_j(a_j + d_j) & \forall j \in FI(i) \end{cases} \quad (1)$$

$$\begin{cases} \gamma_i = T & \text{if } w(k, i) > 0 \text{ or } i \in PO \\ \gamma_i = \min_j(\gamma_j - d_j) & \forall j \in FI(i) \end{cases} \quad (2)$$

where PI is set of all primary inputs and PO is set of all primary outputs. $FI(i)$ and $FO(i)$ represent the incoming and outgoing gates to gate i respectively. Then slack s_i is then calculated by

$$s_i = \gamma_i - a_i \quad (3)$$

A retiming of a circuit G is an integer-valued vertex-labeling r , which represent how many FFs are moved from the outgoing

edges to the incoming edges of each vertex. Thus the number of FFs on edge (i, j) with label r is formulated as follow:

$$w_{i,j} + r_j - r_i$$

Definition 1: Power Slack Curve - Each gate i is given k discrete slack levels, and the power-slack tradeoff is represented by $\{(s_i^1, P(s_i^1)), \dots, (s_i^k, P(s_i^k))\}$. In the Power Slack Curve, each point is connected to its neighboring point(s) by a linear segment.

Based on the relationship between power reduction and slack provided by [15], we assume Power Slack Curve is a convex decreasing function.

Definition 2: Simultaneous Slack Budget and Retiming Problem - Given a directed graph $G = (V, E, d, w)$ representing a synchronous sequential circuit, and period constraint T , we want to find FFs reallocation represented by r , such that the power consumption obtained by slack budgeting is minimized under the period constraint.

According to the above definitions and notations, the simultaneous slack budget and retiming problem can be easily formulated into the following mathematical program.

$$\min \sum_{i \in V} P(s_i) \quad (I)$$

$$\text{s.t. } (1) - (3)$$

$$r_j - r_i \geq -w_{i,j} \quad \forall (i, j) \in E$$

$$s_i \in \{s_i^1, \dots, s_i^k\} \quad \forall i \in V$$

$$a_i \leq T \quad \forall i \in V$$

3. METHODOLOGY

3.1. MILP Formulation

The MILP formulation for retiming synchronous circuits is originally presented in [1] to minimize clock period. The clock period $\Phi(G) \leq T$ if and only if there exists an assignment of real values a_i and an integer value r_i to each vertex $i \in V$ such that the following conditions are satisfied:

$$a_i \geq d_i + s_i \quad \forall i \in V \quad (4)$$

$$a_i \leq T \quad \forall i \in V \quad (5)$$

$$r_i - r_j \leq w_{ij} \quad \forall (i, j) \in E \quad (6)$$

$$a_j \geq a_i + d_i + s_i \quad \text{if } r_i - r_j = w_{ij} \quad (7)$$

Suppose $R_i = r_i + a_i/T$, then $a_i = T \cdot R_i - T \cdot r_i$. The problem can be formulated as (II).

$$\min \sum_{i \in V} P(\bar{s}_i) \quad (II)$$

$$\text{s.t. } \bar{R}_i - \bar{r}_i \geq \bar{s}_i \quad \forall i \in V \quad (IIa)$$

$$\bar{R}_i - \bar{r}_i \leq T \quad \forall i \in V \quad (IIb)$$

$$\bar{r}_j - \bar{r}_i \geq -T \cdot w_{ij} \quad \forall (i, j) \in E \quad (IIc)$$

$$0 \leq \bar{R}_i, \bar{r}_i \leq \bar{N}_{ff} \quad \forall i \in V \quad (IId)$$

$$\bar{s}_i \in \{\bar{s}_i^1, \dots, \bar{s}_i^k\} \quad \forall i \in V \quad (IIe)$$

$$0 \leq \bar{s}_i \leq T \quad \forall i \in V \quad (IIIf)$$

$$\bar{R}_j - \bar{R}_i \geq t_{ij} \quad \forall (i, j) \in E \quad (IIg)$$

$$t_{ij} \geq \bar{s}_j - T \cdot w_{ij} \quad \forall (i, j) \in E \quad (IIh)$$

where $\bar{N}_{ff} = N_{ff} \cdot T$, $\bar{s}_i = d_i + s_i$, $\bar{r}_i = r_i \cdot T$ and $\bar{R}_i = R_i \cdot T$. For each gate i , $\bar{s}_i^j = s_i^j + d_i (j = 1, \dots, k)$.

This problem can be solved by common ILP solver. However, computationally ILP is one of the most difficult combinatorial optimization problems and the runtime is unacceptable even if the problem size is small. In the following subsections, we will explain how to transform this problem to a convex cost dual network flow problem.

3.2. Formulation Simplification

Constraint (IIh) make problem (II) too complex to solve by network flow-based algorithm. First we consider a more simple formulation (III), which removes constraint (IIh). To compensate the lose of accuracy, we add penalty function $P(t_{ij})$ in objective function.

$$\begin{aligned} \min \quad & \sum_{i \in V} P(\bar{s}_i) + \sum_{(i,j) \in E} P(t_{ij}) \\ \text{s.t.} \quad & (IIa) - (IIg) \\ & t_{ij} \geq -c \cdot w_{ij}, \quad \forall (i,j) \in E \end{aligned} \quad (III)$$

where $P(t_{ij}) = P(\bar{s}_j)/k$, and k is a coefficient. Here we set $k = \sum_i (1 - w_{ij})^1$.

Given solution of problem (III) $\bar{s}_i (i = 1, \dots, m)$ and $t_{ij} (\forall (i,j) \in E)$, we propose a heuristic method to generate solution of problem (II).

$$t_{i,j} \geq \bar{s}_j - c \cdot w_{ij} \Rightarrow \bar{s}_j = \min(t_{ij} + c \cdot w_{ij}), \forall i \in FI(j) \quad (8)$$

We denote the \bar{s}_j got in (8) as $\bar{s}_j(\Omega)$ and \bar{s}_j got from problem (III) as $\bar{s}_j(\Theta)$, then we can get \bar{s}_j in problem (II) as follows:

$$\begin{aligned} \bar{s}_j &= \min[\bar{s}_j(\Omega), \bar{s}_j(\Theta)] \\ &= \min[\min(t_{ij} + c \cdot w_{ij}), \bar{s}_j(\Theta)], \quad \forall i \in FI(j) \end{aligned} \quad (9)$$

By now we have build the connection between solution of problem (II) and problem (III). After we calculate the solution of (III), we can then get the solution of (II). In the next subsection, we will prove problem (III) can be transformed to convex cost dual network flow problem.

3.3. Remove Redundant Constraint

In this subsection we will prove that without loss of optimality, problem (III) can remove constraint $\bar{R}_i - \bar{r}_i \leq T$.

Let s_i^* denote the value of s_i for which $P(\bar{s}_i)$ is minimum. In case there are multiple values for which $P(\bar{s}_i)$ is minimum, the minimum value will be chosen. Let us define the function $Q(\bar{s}_i)$ in the following manner:

$$Q(\bar{s}_i) = \begin{cases} P(\bar{s}_i^*) & \text{if } \bar{s}_i \leq s_i^* \\ P(\bar{s}_i) & \text{if } \bar{s}_i > s_i^* \end{cases} \quad (10)$$

¹We suppose for each $(i,j) \in E$, w_{ij} is 0-1 variable.

Now consider the following problem (III'), which replaces (IIa) and (IIb) by $\bar{R}_i - \bar{r}_i = \bar{s}_i$:

$$\begin{aligned} \min \quad & \sum_{i \in V} Q(\bar{s}_i) + \sum_{(i,j) \in E} P(t_{ij}) \\ \text{s.t.} \quad & (IIc) - (IIg) \\ & \bar{R}_i - \bar{r}_i = \bar{s}_i \quad \forall i \in V \\ & t_{ij} \geq -T \cdot w_{ij} \quad \forall (i,j) \in E \end{aligned} \quad (III')$$

Theorem 1: For every optimal solution $(\bar{R}, \bar{r}, \bar{s})$ of problem (III), there is an optimal solution $(\bar{R}, \bar{r}, \hat{s})$ of problem (III'), and the converse also holds.

Proof: Consider an optimal solution $(\bar{R}, \bar{r}, \bar{s})$ of (III), we show how to construct an optimal solution $(\bar{R}, \bar{r}, \hat{s})$ of (III') with the same cost. There are two cases to consider:

Case 1: $\bar{R}_i - \bar{r}_i \geq s_i^*$. It follows from (IIa) and the convexity of $P(\bar{s}_i)$ that $\hat{s}_i = s_i^*$. In this case, we set $\hat{s} = \bar{R}_i - \bar{r}_i$. It follows from (10) that $P(\bar{s}_i) = Q(\hat{s}_i)$.

Case 2: $\bar{R}_i - \bar{r}_i < s_i^*$. Similar to case 1, we can get $\bar{s}_i = \bar{R}_i - \bar{r}_i$. In this case, we set $\hat{s}_i = \bar{R}_i - \bar{r}_i$. It follows from (10) that $P(\bar{s}_i) = Q(\hat{s}_i)$.

Similarly, it can be shown that if $(\hat{R}, \hat{r}, \hat{s})$ is an optimal solution of (III'), then the solution $(\bar{R}, \bar{r}, \bar{s})$ constructed in the following manner is an optimal solution of (III): $\bar{s}_i = \max\{s_i^*, \hat{s}_i\}$. ■

Theorem 2: The constraint $\bar{R}_i - \bar{r}_i \leq T$ in problem (III) can be removed.

Proof: By Theorem 1, we can transform each constraint in (IIa) to an equality constraint. In other words, $\bar{R}_i - \bar{r}_i = \bar{s}_i$. Because constraint (IIb) ($0 \leq \bar{s}_i \leq T$), $\bar{R}_i - \bar{r}_i \leq T$. So we can remove constraint $\bar{R}_i - \bar{r}_i \leq T$. ■

3.4. Transformation to Primal Network Flow Problem

To further simplify problem (III), we transform $G(V, E)$ into $\bar{G}(\bar{V}, \bar{E})$ in such a way that each vertex $i \in V$ is split into two vertex \bar{r}_i and \bar{R}_i . So constraints (IIa) (IIg) and (IIc) can be transformed to the connection relationship in \bar{E} . $\bar{V} = \{\bar{r}_1, \bar{R}_1, \dots, \bar{r}_m, \bar{R}_m\}$. $\bar{E} = \bar{E}_1 \cup \bar{E}_2 \cup \bar{E}_3$, where \bar{E}_1 include edges (\bar{r}_i, \bar{R}_i) , \bar{E}_2 include edges (\bar{R}_i, \bar{R}_j) and edges (\bar{r}_i, \bar{r}_j) belong to \bar{E}_3 . Fig. (2a) illustrates a simple DAG G representing a synchronous sequential circuit, and the transformed DAG \bar{G} of G is illustrated in Fig. (2b).

Now the problem formulation can be simplified as follows:

$$\begin{aligned} \min \quad & \sum_{(i,j) \in \bar{E}} P(s_{ij}) \\ \text{s.t.} \quad & \mu_j - \mu_i \geq s_{ij} \quad \forall (i,j) \in \bar{E} \\ & 0 \leq \mu_i \leq \bar{N}_{ff} \quad \forall i \in \bar{V} \\ & l_{ij} \leq s_{ij} \leq u_{ij} \quad \forall (i,j) \in \bar{E} \end{aligned} \quad \begin{aligned} (IV) \\ (IVa) \\ (IVb) \\ (IVc) \end{aligned}$$

where s_{ij} represents slack assigned to edge from node i to j . For each edge $e(i,j) \in \bar{E}_1$, if $i = \bar{r}_p$ and $j = \bar{R}_p$, then $s_{ij} = \bar{s}_p$, and $l_{ij} = \bar{s}_p^1$ and $u_{ij} = \bar{s}_p^k$. For each edge $e(i,j) \in \bar{E}_2$, $s_{ij} = \bar{s}_j - T \cdot w_{ij}$, then $l_{ij} = \bar{s}_j^1 - T \cdot w_{ij}$ and $u_{ij} = \bar{s}_j^k - T \cdot w_{ij}$.

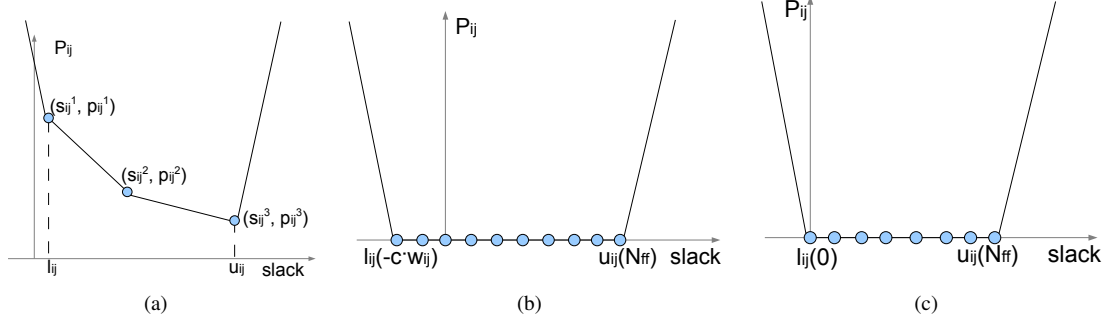


Fig. 3: The Power-Slack Curve of (a)an edge $(i, j) \in E_1 \cup E_2$, here we assume $w_{ij} = 0$; (b)an edge $(i, j) \in E_3$; (c)an edge $(i, j) \in E_4$.

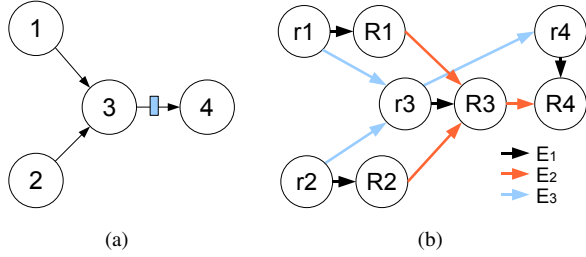


Fig. 2: (a)The DAG G representing a synchronous sequential circuit. (b)The transformed DAG \bar{G} of G .

For each edge $e(i, j) \in E_3$, $l_{ij} = -T \cdot w_{ij}$ and $u_{ij} = \bar{N}_{ff}$. An example Power-Slack Curve of an edge in $E_1 \cup E_2$ and that of an edge in E_3 are illustrated in Fig. (3a) and Fig. (3b), respectively.

We then further eliminate constraints (IVb) and (IVc). First of all, $P(s_{ij})$ can be modified to eliminate the bounds on \bar{s}_i as follows.

$$\bar{P}(s_{ij}) = \begin{cases} P(u_{ij}) + M(s_{ij} - u_{ij}) & \bar{s}_{ij} > u_{ij} \\ P(s_{ij}) & 0 \leq \bar{s}_{ij} \leq T \\ P(l_{ij}) - M(s_{ij} - l_{ij}) & \bar{s}_{ij} < l_{ij} \end{cases} \quad (11)$$

where M is a sufficiently large number such that $\bar{P}(s_{ij})$ is still a convex function.

Similarly, the bounds on μ_i can also be eliminated by adding into objective a convex cost function $B(\mu_i)$ defined as follows.

$$B(\mu_i) = \begin{cases} M \cdot (\mu_i - \bar{N}_{ff}) & \text{if } \mu_i > \bar{N}_{ff} \\ 0 & \text{if } 0 \leq \mu_i \leq \bar{N}_{ff} \\ -M \cdot \mu_i & \text{if } \mu_i < 0 \end{cases} \quad (12)$$

After the above simplifications, problem (IV) can be transformed to problem (V):

$$\begin{aligned} \min \quad & \sum_{(i,j) \in \bar{E}} \bar{P}(s_{ij}) + \sum_{i \in \bar{V}} B(\mu_i) \\ \text{s.t.} \quad & \mu_j - \mu_i \geq s_{ij} \quad \forall (i, j) \in \bar{E} \end{aligned} \quad (V)$$

3.5. Problem Transformation by Lagrangian Relaxation

Using Lagrangian relaxation to eliminate constraint in problem (V), get the Lagrangian sub-problem:

$$\begin{aligned} L(\vec{x}) = & \sum_{e(i,j) \in \bar{E}} \bar{P}(s_{ij}) + \sum_{i \in \bar{V}} B(\mu_i) \\ & - \sum_{e(i,j) \in \bar{E}} (\mu_j - \mu_i - s_{ij})x_{ij} \end{aligned} \quad (13)$$

It is easy to show that

$$\sum_{e(i,j) \in \bar{E}} (u_i - u_j)x_{ij} = \sum_{i \in \bar{V}} x_{0i} \times \mu_i \quad (14)$$

where

$$x_{0i} = \sum_{j: e(i,j) \in \bar{E}} x_{ij} - \sum_{j: e(j,i) \in \bar{E}} x_{ji}, \forall i \in V \quad (15)$$

Lagrangian subproblem (13) can be restated as follows:

$$L(\vec{x}) = \min \sum_{e(i,j) \in \bar{E}} [P(s_{ij}) + x_{ij}s_{ij}] + \sum_{i \in \bar{V}} [B(\mu_i) + x_{0i}\mu_i] \quad (16)$$

A start node v_0 is added to \bar{V} , v_0 interconnects all other nodes in \bar{V} . We set $s_{0i} = \mu_i, l_{0i} = 0, u_{0i} = \bar{N}_{ff}$. So $V = \{v_0\} \cup \bar{V}$. The new edges are denoted as E_4 , $E = \bar{E} \cup E_4$. The Power-Slack curve of an edge $(i, j) \in E_4$ is illustrated in Fig. (3c). So we can transform $L(\vec{x})$ as formulation (17).

$$\begin{aligned} L(\vec{x}) = \min \quad & \sum_{e(i,j) \in E} [P_{ij}(s_{ij}) + x_{ij}s_{ij}] \\ \text{s.t.} \quad & \sum_{j: e(i,j) \in E} x_{ij} - \sum_{j: e(j,i) \in E} x_{ji} = 0 \quad \forall i \in V \\ & x_{ij} \geq 0 \quad \forall (i, j) \in E_1 \cup E_2 \cup E_3 \end{aligned} \quad (17)$$

3.6. Convex Cost-scaling Approach

We define function $H_{ij}(x_{ij})$ for each $e(i, j) \in E$ as follows:

$$H_{ij}(x_{ij}) = \min_{s_{ij}} \{P_{ij}(s_{ij}) + x_{ij}s_{ij}\} \quad (18)$$

For the $e(i, j) \in E_1$, because the function $H_{ij}(x_{ij})$ is a piecewise linear concave function of x_{ij} , and $\forall e(i, j) \in E_1$, then $H_{ij}(x_{ij})$ is described in the following manner [19]:

$$H_{ij}(x_{ij}) = \begin{cases} P_{ij}(s_{ij}^k) + s_{ij}^k x_{ij} & 0 \leq x_{ij} \leq b_{ij}(k) \\ \dots \\ P_{ij}(s_{ij}^q) + s_{ij}^q x_{ij} & b_{ij}(q+1) \leq x_{ij} \leq b_{ij}(q) \\ \dots \\ P_{ij}(s_{ij}^1) + s_{ij}^1 x_{ij} & k \leq x_{ij} \end{cases}$$

where $b_{ij}(q) = \frac{P_{ij}(s_{ij}^{q-1}) - P_{ij}(s_{ij}^q)}{s_{ij}^{q-1} - s_{ij}^q}$.

For the $e(i, j) \in E_1$, similar to E_2 , then $H_{ij}(x_{ij}) =$

$$H_{ij}(x_{ij}) = \begin{cases} P_{ij}(t_{ij}^k) + t_{ij}^k x_{ij} & 0 \leq x_{ij} \leq b_{ij}(k) \\ \dots \\ P_{ij}(t_{ij}^q) + t_{ij}^q x_{ij} & b_{ij}(q+1) \leq x_{ij} \leq b_{ij}(q) \\ \dots \\ P_{ij}(t_{ij}^1) + t_{ij}^1 x_{ij} & k \leq x_{ij} \end{cases}$$

where $b_{ij}(q) = \frac{P_{ij}(t_{ij}^{q-1}) - P_{ij}(t_{ij}^q)}{t_{ij}^{q-1} - t_{ij}^q}$, and $t_{ij}^q = s_{ij}^q - T \cdot w_{ij}$.

For the $e(i, j) \in E_3$, because $P_{ij}(s_{ij}) = 0$,

$$H_{ij}(x_{ij}) = \min_{s_{ij}} (s_{ij} x_{ij}) = -T \cdot w_{ij} \cdot x_{ij}, x_{ij} \geq 0$$

For the $e(i, j) \in E_4$, the variable $x_{i,j}$ is not a Lagrangian multiplier, and it is bounded by $-M \leq x_{ij} \leq M$.

$$H_{ij}(x_{ij}) = \begin{cases} 0 & 0 \leq x_{ij} \leq M \\ \bar{N}_{ff} \cdot x_{ij} & -M \leq x_{ij} \leq 0 \end{cases}$$

Note that these functions $H_{ij}(x_{ij})$ are all concave. We define $C_{ij}(x_{ij}) = -H_{ij}(x_{ij})$, so that $C_{ij}(x_{ij})$ is a piecewise linear convex function. Then we can subsequently propose problem (VI) as follows:

$$\begin{aligned} L(\vec{x}) &= \min \sum_{e(i,j) \in E} C_{ij}(x_{ij}) \\ \text{s.t.} \quad & \sum_{j:e(i,j) \in E} x_{ij} - \sum_{j:e(j,i) \in E} x_{ji} = 0 \quad \forall i \in V \\ & 0 \leq x_{ij} \leq M \quad \forall (i, j) \in E_1 \cup E_2 \cup E_3 \\ & -M \leq x_{ij} \leq M \quad \forall (i, j) \in E_4 \end{aligned} \quad (VI)$$

To transform the problem into a minimum cost flow problem, we construct an expanded network $G' = (V', E')$. There are four kinds of edges to consider:

- $e(i, j)$ in E_1 : we introduce k edges in G' , and the costs of these edges are: $-s_{ij}^k, -s_{ij}^{k-1}, \dots, -s_{ij}^1$; upper capacities: $b_{ij}(k), b_{ij}(k-1) - b_{ij}(k), b_{ij}(k-2) - b_{ij}(k-1), \dots, M - b_{ij}(2)$, where M is a huge coefficient; lower capacities are all 0.
- $e(i, j)$ in E_2 : we introduce k edges in G' , and the costs of these edges are: $-t_{ij}^k, -t_{ij}^{k-1}, \dots, -t_{ij}^1$; upper capacities: $b_{ij}(k), b_{ij}(k-1) - b_{ij}(k), b_{ij}(k-2) - b_{ij}(k-1), \dots, M - b_{ij}(2)$, where M is a huge coefficient; lower capacities are all 0.
- $e(i, j)$ in E_3 : cost, lower and upper capacity is $(c \cdot w_{ij}, 0, M)$.

TABLE I: Characteristics of Test Cases

Case Name	Gate #	Edges #	Max Output	Max Inputs	Tmin
s27.test	11	19	4	2	20
s208.1.test	105	182	19	4	28
s298.test	120	250	13	6	24
s382.test	159	312	21	6	44
s386.test	160	354	36	7	64
s344.test	161	280	12	11	46
s349.test	162	284	12	11	46
s444.test	182	358	22	6	46
s526.test	194	451	13	6	42
s526n.test	195	451	13	6	42
s510.test	212	431	28	7	42
s420.1.test	219	384	31	4	50
s832.test	288	788	107	19	98
s820.test	290	776	106	19	92
s641.test	380	563	35	24	238
s713.test	394	614	35	23	262
s838.1.test	447	788	55	4	80
s1238.test	509	1055	192	14	110
s1488.test	654	1406	56	19	166

- $e(i, j)$ in E_4 : two edges are introduced in G' , one with cost, lower and upper capacity as $(\bar{N}_{ff}, -M, 0)$, another is $(0, 0, M)$.

Using the cost-scaling algorithm [24], we can solve the minimum cost flow problem in G' . For the given optimal flow x^* , we construct residual network $G(x^*)$ and solve a shortest path problem to determine shortest path distance $d(i)$ from node s to every other node. By implying that $\mu(i) = d(i)$ and $s_{ij} = \mu(i) - \mu(j)$ for each $e(i, j) \in E_1 \cup E_2$, we can finally solve problem (III).

4. EXPERIMENTAL RESULTS

We implemented our algorithm in the C++ programming language and executed on a Linux machine with eight 3.0GHz CPU and 6GB Memory. 19 cases from the ISCAS89 benchmarks are tested, and the name, number of gates, number of signal passes, the maximum number of gate output/inputs, and the minimum period for each case are given in Table I. We used four discrete slack levels for each gate as $\{0, 10, 20, 33\}$. Energy consumption of the gates with slack level scaling were found from model in [15].

In the experiments, a min-period retiming algorithm [4] is first employed to generate the minimum clock period T , which is listed in the 2nd column of TABLE II. Liu et al.'s [18] algorithm was implemented for comparison. Note that algorithm in [18] can not directly solve discrete slack budgeting problem, because if sensitive transitive closure graph is used, the timing constraints might be violated after slack budgeting [8]. Therefore we use a transitive closure graph instead of sensitive transitive closure graph here. To evaluate the accuracy of our algorithm, the ILP for achieving the optimal solution were also implemented using an open source ILP solver CBC [25].

Table II shows comparisons among optimal ILP, algorithm in [18] and our algorithm. The column Power Consumption gives actual power consumption of each circuit and less value means more power can be reduced. Comparing with optimal

TABLE II: Comparisons with Optimal ILP and Previous Work [18]

Benchmark	T	Power Consumption			Total Slacks			Runtime(s)		
		Optimal ILP	[18]	ours	Optimal ILP	[18]	ours	Optimal ILP	[18]	ours
s27.test	20	800	824	850	40	40	30	0.02	0.0	0.0
s208.1.test	28	3542	9118	4772	1770	290	1988	0.39	0.44	0.06
s298.test	24	6498	8888	8010	1330	660	1240	0.78	0.69	0.07
s382.test	44	6456	9038	9958	3011	2071	1895	>1000	10.56	0.12
s386.test	64	8836	12870	9564	2484	807	2324	4.58	1.03	0.1
s344.test	46	9876	11848	9894	1855	1064	1760	0.82	2.53	0.09
s349.test	46	9938	12472	9894	1852	912	1780	0.79	4.49	0.11
s444.test	46	8938	14032	11884	2962	1025	1939	>1000	12.04	0.12
s526.test	42	7602	14106	11498	3626	1307	2356	42.57	1.67	0.17
s526n.test	42	7752	11734	11548	3616	2089	2366	30.32	4.72	0.17
s510.test	42	13976	17492	14846	2237	937	2040	>1000	1.62	0.17
s420.1.test	50	4574	17920	9224	5906	1050	4466	1.29	16.91	0.14
s832.test	98	13652	14518	16274	5175	4525	4171	71.96	151.26	0.24
s820.test	92	13552	17694	16448	5261	3493	4103	68.98	13.18	0.25
s641.test	238	13334	20408	14424	7925	6067	7604	2.24	92.97	0.26
s713.test	262	13018	21228	14322	8522	6363	8112	2.27	121.1	0.27
s838.1.test	80	6004	18898	17556	14048	9016	9912	1.48	256.9	0.4
s1238.test	110	6096	10444	8208	16764	14635	15792	0.23	448.6	0.34
s1488.test	166	21292	23799	27836	15313	14791	13024	>1000	670.7	0.53
Avg	-	9249.3	14070	11947.9	5457.7	3744.3	4573.8	-	95.3	0.19
Diff	-	1	+52%	+29%	1	-31%	-16%	-	1	0.002

solution, our algorithm increases 29% power consumption while [18] increases 52%. Column Total Slack gives the sum of each gate's slack. Comparing with optimal solution, our algorithm loses 16% of slacks while [18] loses 31%. Note that power consumption is not proportional to the slack amount. As for benchmark s27.test, [18] and optimal ILP get equal slack amount, but their power consumption is different. Column Runtime compares the run time of each algorithm. From the results we can find that although optimal ILP can get optimal solution, its runtime sometimes is unacceptable. Comparing with [18], our algorithm can not only generate better design results, but also get nearly 500× speedup.

5. CONCLUSION

In this paper we have showed that the retiming and slack budgeting problem can be simultaneously solved by formulating the problem to a convex cost dual network flow problem. Both the theoretical analysis and experimental results show the efficiency of our approach which can not only reduce power consumption but also speedup previous work.

REFERENCES

- [1] C. E. Leiserson and J. B. Saxe, "Retiming synchronous circuitry," *Algorithmica*, vol. 6, pp. 5–35, 1991.
- [2] N. Maheshwari and S. Sapatnekar, "Efficient retiming of large circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 6, pp. 74–83, 1998.
- [3] H. Zhou, "Deriving a new efficient algorithm for min-period retiming," in *ACM/IEEE Asia and South Pacific Design Automation Conference (ASPDAC)*, 2005, pp. 990–993.
- [4] —, "A new efficient retiming algorithm derived by formal manipulation," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 13, no. 1, pp. 1–19, 2008.
- [5] C. Lin and H. Zhou, "An efficient retiming algorithm under setup and hold constraints," in *ACM/IEEE Design Automation Conference (DAC)*, 2006, pp. 945–950.
- [6] J. Wang and H. Zhou, "An efficient incremental algorithm for min-area retiming," in *ACM/IEEE Design Automation Conference (DAC)*, 2008, pp. 528–533.
- [7] R. Nair, C. L. Berman, P. S. Hauge, and E. J. Yoffa, "Generation of performance constraints for layout," *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 8, pp. 860–874, 1989.
- [8] D.-S. Chen and M. Sarrafzadeh, "An exact algorithm for low power library-specific gate re-sizing," in *ACM/IEEE Design Automation Conference (DAC)*, 1996, pp. 783–788.
- [9] C. Chen, X. Yang, and M. Sarrafzadeh, "Predicting potential performance for digital circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2002.
- [10] S. Ghiasi, E. Bozorgzadeh, S. Choudhuri, and M. Sarrafzadeh, "A unified theory of timing budget management," in *ACM/IEEE International Conference on Computer Aided Design (ICCAD)*, 2004, pp. 653–659.
- [11] —, "A unified theory of timing budget management," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 2364–2375, 2006.
- [12] D. Nguyen, A. Davare, M. Orshansky, D. Chinnery, B. Thompson, and K. Keutzer, "Minimization of dynamic and static power through joint assignment of threshold voltages and sizing optimization," in *IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, 2003, pp. 158–163.
- [13] A. srivastava, D. Sylvester, and D. Blaauw, "Power minimization using simultaneous gate sizing dual-vdd and dual-vth assignment," in *ACM/IEEE Design Automation Conference (DAC)*, 2004, pp. 783–787.
- [14] S. Kulkarni, A. Srivastava, and D. Sylvester, "A new algorithm for improved vdd assignment in low power dual vdd systems," in *IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, 2004, pp. 200–205.
- [15] X. Qiu, Y. Ma, X. He, and X. Hong, "Iposa: A novel slack distribution algorithm for interconnect power optimization," in *International Symposium on Quality of Electronic Design (ISQED)*, 2008, pp. 873–876.
- [16] Y. Hu, Y. Lin, L. He, and T. Tuan, "Simultaneous time slack budgeting and retiming for dual-vdd fpga power reduction," in *ACM/IEEE Design Automation Conference (DAC)*, 2006, pp. 478–483.
- [17] C. Lin, A. Xie, and H. Zhou, "Design closure driven delay relaxation

based on convex cost network flow,” in *the conference on Design, Automation and Test in Europe (DATE)*, 2007, pp. 63–68.

- [18] S. Liu, Y. Ma, X. Hong, and Y. Wang, “Simultaneous slack budgeting and retiming for synchronous circuits optimization,” in *ACM/IEEE Asia and South Pacific Design Automation Conference (ASPDAC)*, 2010.
- [19] R. K. Ahuja, D. S. Hochbaum, and J. B. Orlin, “Solving the convex cost integer dual network flow problem,” *Manage. Sci.*, vol. 49, no. 7, pp. 950–964, 2003.
- [20] R. Chen and H. Zhou, “Efficient algorithms for buffer insertion in general circuits based on network flow,” in *ACM/IEEE International Conference on Computer Aided Design (ICCAD)*, 2005, pp. 322–326.
- [21] Q. Ma and F. Young, “Network flow-based power optimization under timing constraints in msv-driven floorplanning,” in *ACM/IEEE International Conference on Computer Aided Design (ICCAD)*, 2008, pp. 1–8.
- [22] B. Yu, S. Dong, S. Goto, and S. Chen, “Voltage-island driven floorplanning considering level-shifter positions,” in *ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2009, pp. 51–56.
- [23] C. Lin and H. Zhou, “Clock skew scheduling with delay padding for prescribed skew domains,” in *ACM/IEEE Asia and South Pacific Design Automation Conference (ASPDAC)*, 2007, pp. 541–546.
- [24] R.K.Ahuja, T.L.Magnanti, and J.B.Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall/Pearson, 2005.
- [25] [Online]. Available: <http://www.coin-or.org/projects/Cbc.xml>